La matematica di Google

Federica Galluzzi

30 marzo 2016

Torino, Dipartimento di Matematica

Motori di ricerca

- navigano nel web e individuano tutte le pagine ad accesso pubblico
- indicizzano i dati ottenuti per poter individuare efficientemente le pagine (parole o frasi chiave)
- rispondono alle richieste degli utenti ordinando le pagine per importanza (Page Rank)

Come classificare le pagine per importanza?

Nelle situazioni sociali comuni, una persona è importante (popolare) se altre persone parlano di lei.
Diventa ancora piu' popolare se

- persone popolari parlano di lei
- le persone che parlano di lei non parlano anche di molti altri (in questo caso l'importanza viene divisa tra tutti)

Google la matematica di google Federica

Tutti Notizie Immagini Video Maps Altro Strumenti di ricerca

Circa 628.000 risultati (0,71 secondi)

Math Inside [#1]: La matematica di Google - Maddmaths!
maddmaths.simai.eu/divulgazione/math-inside-1-la-**matematica**-**di**-**google**/
08 mar 2013 - Che c'entra **Google** con **la matematica**? Ce lo racconta Roberto Natalini, Dirigente **di** Ricerca dell'IAC-CNR. Comincia la serie "Math Inside", ...

La matematica di Google | Math is in the air
www.mathisintheair.org/.../la-**matematica**-alla-base-**di**-**google**-lalgoritmo-...
27 set 2015 - In questo post ci "limiteremo" ad analizzare l'aspetto distintivo, almeno nella fase iniziale, alla base **di Google** Search (il PageRank) e i suoi ...

[PDF] "Page ranking" di Google. - Università degli Studi dell'Aquila
univaq.it/~leonetti/lauree_sc/**google**.pdf
Google: **la matematica** per essere migliori. Antonio Cicone ... Queste note nascono dalla lezione **di** Antonio Cicone svolta il 22 novembre 2010 agli studenti del ...

Seminario: “Matematica e Internet : il problema di Google ...
www.iit.cnr.it/it/node/6984
Internet costituisce una sorgente **di** problemi **matematici di** particolare interesse teorico e applicativo. Tra questi il problema del Page ranking **di Google**, cioè il ...

[PDF] Matematica del web: il problema di Google - Dipartimento …
www.dm.unipi.it/~bini/Didattica/Slides/**google**.pdf
di DA Bini - 2006 - Articoli correlati
Outline. **Matematica** e applicazioni. Il problema del Page Rank. Sistemi Lineari. Il navigatore virtuale. **Matematica** del web: il problema **di Google**. Dario A. Bini.

Googol - Wikipedia
https://it.wikipedia.org/wiki/Googol

Il Page Rank di Google

- Non si basa sulle richieste degli utenti (page Rank statico)
- Non si basa sul contenuto delle pagine
- Conta e pesa i link che puntano alle pagine (backlinks)

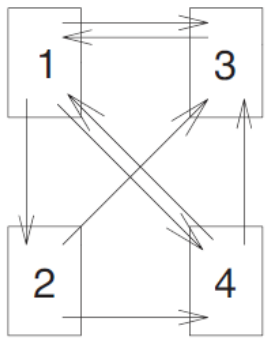

FIG. 2.1. *An example of a web with only four pages. An arrow from page A to page B indicates a link from page A to page B.*

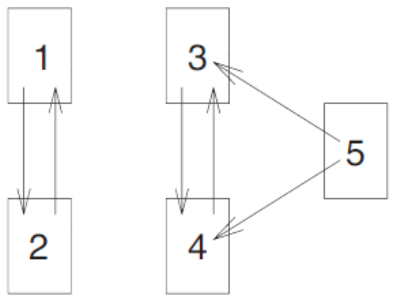

FIG. 2.2. *A web of five pages, consisting of two disconnected "subwebs" W_1 (pages 1 and 2) and W_2 (pages 3, 4, 5).*

Catene di Markov

Una catena di Markov a stati discreti è un processo stocastico $\{X_t\}$ discreto (indicizzato cioè da una variabile t discreta, $t \in T = \mathbb{N}$) in cui lo stato futuro,condizionatamente a quello presente, è indipendente dagli stati passati. La probabilità condizionata di un evento futuro, dati lo stato attuale e tutti gli eventi passati, dipende esclusivamente dallo stato attuale del processo e non dai precedenti. La probabilità condizionata

$$P(X_{t+1} = j | X_t = i) = p_{i,j}(t)$$

è la probabilità di transizione al tempo t dallo stato i allo stato j. La matrice $(p_{i,j}(t))$ è una matrice stocastica.

- Una catena di Markov si dice *omogenea* se per ogni coppia di stati i, j si ha che $P(X_{t+1} = j | X_t = i)$ è indipendente da t, quindi $p_{i,j}(t) = p_{i,j}$
- Una catena di Markov si dice *irriducibile* se la probabilità di passare da uno stato all'altro dell'insieme degli stati è sempre maggiore di zero.
- Il teorema di esistenza e unicità afferma che data una catena di Markov omogenea ed irriducibile, esiste un'unica distribuzione stazionaria (cioè costante nel tempo).
- Il vettore PageRank è il vettore corrispondente alla distribuzione stazionaria della catena. Esiste ed è unico. Può essere trovato con un processo iterativo basato sul metodo delle potenze.

La deformazione della matrice A con parametro $m = 0.15$ dà un buon bilanciamento tra efficacia ed efficienza dell'algoritmo di calcolo tramite iterazioni successive

- Efficacia: fornisce spesso una soluzione molto vicina a quella esatta
- Efficienza: costa meno di un algoritmo esatto (il calcolo esatto dell'autovettore)

Random Web-surfer model

Navigatore casuale che si muove su una rete di n nodi senza pozzi. Inizia da una pagina random. Si muove da una pagina all'altra con questa procedura: se la pagina su cui si trova ha r outgoing links

- può scegliere di spostarsi su uno degli r links con probabilità uniforme $\frac{1-m}{r}$
- OR può saltare su una qualsiasi altra pagina con probabilità $\frac{m}{n}$ (N.B. $r\frac{1-m}{r} + n\frac{m}{n} = 1$)

La componente x_j del vettore di PageRank normalizzato $\mathbf{x}$ è la frazione di tempo che il surfer spende rimanendo sulla pagina j

Struttura del web, da "Graph Structure in the Web"

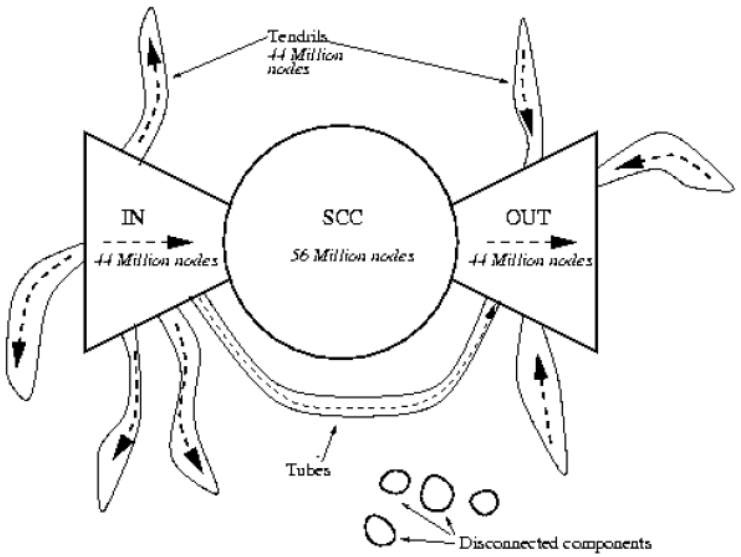

Figure 9: Connectivity of the web: one can pass from any node of IN through SCC to any node of OUT. Hanging off IN and OUT are TENDRILS containing nodes that are reachable from portions of IN, or that can reach portions of OUT, without passage through SCC. It is possible for a TENDRIL hanging off from IN to be hooked into a TENDRIL leading into OUT, forming a TUBE -- a passage from a portion of IN to a portion of OUT without touching SCC.

We now give a more detailed description of the structure in Figure 9. The sizes of the various components are as follows:

Region	SCC	IN	OUT	TENDRILS	DISC.	Total
Size	56,463,993	43,343,168	43,166,185	43,797,944	16,777,756	203,549,046

Bibliografia

- S. Brin, L. Page, "The anatomy of a large-scale hypertextual Web search engine" . Computer Networks and ISDN Systems 30: 107-117. (1998)
- A. Broder et al, "Graph Structure in the Web". www.cis.upenn.edu/ mkearns/teaching/NetworkedLife/broder.pdf
- K. Bryan , T. Leise "The 25,000,000,000 eigenvector: the linear algebra behind Google" SIAM Rev., 48(3), 569-581 (2006)
- http://news.netcraft.com/archives/category/web-server-survey/